序号	技术参数指标要求
1	计算服务器
★1.1	服务器数量不少于 2 台
●1.2	标准机架式服务器
★1.3	CPU：单台实配≥2 颗国产自主可控 ARM CPU，支持超线程，每颗 CPU 核心数≥64，主频≥2.8GHz
★1.4	NPU/GPU：≥2 块国产自主可控 AI 卡，单卡 HBM 内存≥32GB，单卡 FP16 半精度算力≥280TFLOPS，单卡 INT8 整型算力≥560TOPS，单卡功耗≤350W
●1.5	提供 AI 卡厂商针对本项目的授权，提供设备厂商针对本项目的驱动、软件、工具下载授权
★1.6	内存：≥32 根 32GB DDR5 4800MT/s ECC 内存条
★1.7	硬盘：≥2 块 480GB SATA SSD 企业级系统盘；≥6 块 3.84TB SATA SSD 企业级数据盘。模型及数据保存盘：整套系统统一配置一个 5 盘位 USB 硬盘阵列，支持 USB3.0 接口，支持 8 种 RAID 配置，最大容量支持 120TB，实配≥2 块 20TB 企业级硬盘。
▲1.8	网络：≥1 个 100GE 光口，≥2 个 25GE 光口，≥2 个 GE 电口
▲1.9	阵列卡：配置独立 RAID 卡,支持 RAID0,1,5,6,10,50,60
●1.10	电源：≥2 个白金高效率电源，支持热插拔及 1+1 冗余，单电源功率≥2000W，并提供配套的电源连接线
●1.11	管理系统：配置独立千兆管理网口，禁止与业务口混用；配置 IPMI、远程监控图形界面，可远程对服务器控制；提供资源状态监控与外部编程接口，可对第三方管理软件提供当前资源状态信息
●1.12	服务器管理系统配置国产管理芯片，并提供原厂商证明材料
▲1.13	提供国产自主研发的深度匹配所投 CPU 的数学库、编译器、并行通信库软件，并提供 CPU 芯片厂商官方软件社区链接和截图证明
▲1.14	AI 开发框架：所投 AI 卡支持 Pytorch、MindSpore、TensorFlow 等主流 AI 开发框架，并提供 AI 芯片厂商官方软件社区链接和截图证明，给出对上述框架支持的版本以及后续升级版本的支持能力说明
▲1.15	AI 驱动：提供针对所投 AI 卡的国产自主可控的驱动软件，对上支持多种

	AI 框架，对下服务 AI 芯片与编程，提供高效易用的编程接口，并提供 AI 芯片厂商官方软件社区链接和截图证明
★1.16	推理引擎： 1.提供针对所投 AI 卡的国产自主可控的推理引擎，支持推理场景下的运行加速、调试调优、快速迁移部署 2. 能够运行全参数（671B） DeepSeek V3 R1。精度大于 int4，单请求的输出性能不低于 15 tps，2 个 请求下综合的 tps 不低于 25 tps，提供可复现的测试报告作为证明材料 3.支持 prefix cache 功能，对已经缓存的请求能够快速响应